AFML-TR-68-367

DOCUMENT RETRIEVAL SYSTEM OPERATIONS INCLUDING THE USE OF MICROFICHE AND THE FORMULATION OF A COMPUTER AIDED INDEXING CONCEPT

F. L. SCHEFFLER

R. B. SMITH

University of Dayton Research Institute

Dayton, Ohio

TECHNICAL REPORT AFML-TR-68-367

FEBRUARY 1969

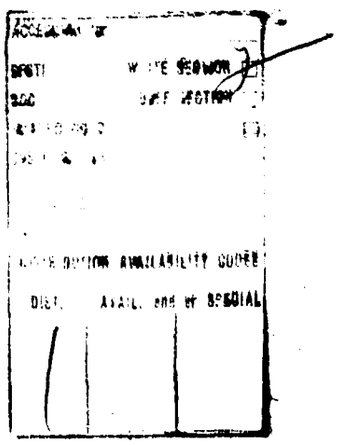
AIR FORCE MATERIALS LABORATORY
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO

50

NOTICES

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This document has been approved for public release and sale; its distribution is unlimited.



Copies of this report should not be returned unless return is required by security considerations, contractual obligations, or notice on a specific document.

700 - April 1969 - CO455 - 72-1593

DOCUMENT RETRIEVAL SYSTEM OPERATIONS INCLUDING THE USE OF MICROFICHE AND THE FORMULATION OF A COMPUTER AIDED INDEXING CONCEPT

F. L. SCHEFFLER

R. B. SMITH

FOREWORD

This report was prepared by the University of Dayton Research Institute, Dayton, Ohio under Air Force Contract AF 33(615)-3389. The work described herein was accomplished under project 7381 "Materials Application" and Task No. 738103 "Materials Information Development Collection and Processing." The effort was administered under the direction of the Materials Information Branch, Materials Support Division, Air Force Materials Laboratory with H. B. Thompson, MAAM, as project monitor.

This is a summary technical report and covers the work accomplished from 1 December 1967 through 30 November 1968.

The authors acknowledge the efforts and contributions of the Project Supervisor, Edward A. Janning, their co-workers Eugene R. Egan, Howard S. Schumacher, Jr. and Linda C. Haley and the part-time indexing personnel, Charles W. Minnis, George J. Pacinda, Michael J. Luthman, Jacqueline F. March, and Robert E. Bauerle. Mr. Paul Evans and Mr. Roger Buening contributed significantly to the computer aided indexing project by providing programming support. Mr. Edward L. Horne of the Materials Information Branch, AFML coordinated the CAS SDI Program.

This report was submitted by the authors December 1968.

This technical report has been reviewed and is approved.

Edward Dugger
Chief, Materials Information Branch
Materials Support Division
Air Force Materials Laboratory

ABSTRACT

Additional part-time indexers were trained according to a previously established training program. Many documents are now being received in microfiche form. Several makes of microfiche readers were evaluated, particularly with regard to their use by indexers. One make offered the most advantages, and a number of these readers were purchased. A method of producing typed hard copy abstracts from microfiche was found. The number of search requests increased with requests both from outside organizations and from the AFML accounting for the increase. The thesaurus was updated and a separate section containing only metallic materials terminology made the thesaurus easier to use.

A computer aided indexing concept was formulated which is based on analysis of technical text. Text words are matched against files of nontechnical words, technical words and bound terms. Words not recognized are presented for human intellectual decisions which are subsequently incorporated into the system. The concept included a self-contained document analysis, storage and retrieval system with which requestors could interact by means of remote access terminals. Remote communication capability with the computer was accomplished, some programs were prepared and file structures based on word length and alphabetization of the first two characters were designed.

This abstract has been approved for public release and sale; its distribution is unlimited.

TABLE OF CONTENTS

Section I

INTRODUCTION

The Information Systems Section of the University of Dayton Research Institute has established and presently maintains and operates a document retrieval system in support of the Aerospace Materials Information Center (AMIC). The document retrieval system operated by the University of Dayton contains approximately 37,000 documents concerning materials research and development with new accessions being made continually. The establishment, modification and operation of the document retrieval system are described in the following reports: RTD-TDR-63-4263 (AD 428 423)[1], AFML-TR-65-20 (AD 613 301)[2], AFML-TR-66-36 (AD 633 614)[3], AFML-TR-66-391 (AD 651 039)[4], and AFML-TR-679 (AD 666 462)[5]. The present report describes the work performed from December 1967 to December 1968.

Additional part-time indexers were employed and the indexer training program described in Reference 4 was applied to their training as indexers. Their performance was evaluated as described in Reference 5, and a comparison of their performance was made with that of previous trainees who had undergone the training program. A major consideration was the acquisition of the majority of documents for the document retrieval system in microfiche form. In conjunction with the change-over from hard copy documents to documents in microfiche form, certain considerations were necessary. Several types of portable microfiche readers were evaluated specifically with regard to their suitability for use by indexers. Problems in handling microfiche were studied. Several means of obtaining a typed abstract for each document for inclusion into the abstract card reference files were evaluated. The University's previous experience with microform records was valuable in dealing with documents in microfiche form.

Statistical data indicate an increase both in document acquisition and an increase in the number of searches performed over the preceding year. Handbooks are now being indexed and included in the system. The thesaurus was updated, and the feasibility of separating the thesaurus terminology was investigated. The Chemical Abstracts Services SDI programs were continued. Efforts are continuing toward developing a system for indexing and retrieval of information science documents. The basic concept of a computer aided indexing scheme was developed. With this scheme, technical text is computer analyzed and the technical concepts are derived therefrom. The idea included a self-contained document analysis, storage and retrieval system by which requestors could interact directly with the computer system via remote terminals.

Section II

TRAINING OF PART-TIME INDEXERS

An effective indexer training program has been in use at the University for about two years. During the past year, five new part-time indexers have been trained by means of this training program. Four of these trainees were undergraduate students - two in chemistry, one in physics and one in chemical engineering. The other trainee was a graduate chemist who works part-time as an instructor in the Chemistry Department. The training process used is described in References 4 and 5.

An indexer performance evaluation was carried out for the new indexers at the end of their training period. A comparison of the results from all the indexers who have participated in the training program is presented in Table I. The results indicate that there is some difference between the trainees in 1967 and 1968. The percentage of essential terms and total terms is slightly higher for the trainees in 1968. The number of misleading terms employed by trainees in 1968 is significantly lower than occurred in earlier years. This reduction in misleading terms can be attributed in part to an improvement in the administration of the training program, but it is believed to be due primarily to improvements in the thesaurus. The average indexing time for trainees in 1967 and 1968 was not significantly different.

The indexer training program continues to be an effective means of training new indexers. The close interaction between the trainee and his supervisor, the exposure of the trainees to other aspects of the system's operations and the active encouragement of suggestions from the trainees are considered to contribute to a professional environment which is responsible for the rapid development of indexer trainees into competent indexers.

TABLE I. COMPARISON OF INDEXER TRAINEES WITH EXPERIENCED INDEXERS

Indexer	Avg. No. Tot. Terms Used In Indexg.	Avg. No. Terms Judged Ess.	% Com. Total Terms (FLS 1966)	% Ess. Included in Indxg. (FLS 1966)	% Avg. Misleading Terms Used in Indxg. Based on FLS Ess. Terms	Avg. Time for Indxg. (minutes)
FLS 66[1]	15.5	7.5	100%	100%		25.4
FLS 67[1]	18.6	8.0	77.5%	92.0%		14.3
LCH 67[1]	20.6	10.3	70.0%	86.2%	0.0%	14.5
JFM 68[2]	27.7		67.7%	85.8%	3.1%	35.6
REB 68[3]	24.3		64.6%	85.4%	2.2%	27.1
GJP 68[3]	16.6		54.8%	75.1%	1.9%	26.9
MJL 68[3]	17.7		64.0%	83.9%	1.1%	20.6
CWM 68[3]	23.7		62.9%	84.7%	0.4%	21.5
PAT 67[4]	22.0		65.7%	82.0%	4.6%	24.0
EJF 67[4]	20.7		62.5%	82.6%	1.1%	17.4
LCS 66[3]	15.4		60.0%	82.0%	9.7%	81.6*
TRD 66[3]	12.2		54.0%	71.1%	8.8%	71.0*

1. Experienced Indexer
2. Part-time Professional
3. Undergraduate Student
4. Graduate Student

* The evaluation documents were indexed by the students of 1966 before systematic procedures had been established. Time required for referring to reference materials is included. In subsequent years the reference referral time was not included as indexing time.

Section III

DOCUMENTS IN MICROFICHE FORM

In recent years, microfiche has received increasing consideration as a means of document storage and retrieval. A proposal has been made for microfiche in libraries.[6] Microfiche has already been implemented by certain Federal Government agencies. The Atomic Energy Commission (AEC), the National Aeronautics and Space Administration (NASA) and the Department of Defense (DOD) have standardized in the use of the microfiche form. These three agencies are collectively responsible for publishing over 90 percent of the scientific and technical reports sponsored by the U. S. Government.[7] The standard microfiche form accepted by these agencies is 105 mm x 148 mm (4" x 6") and contains 60 page images per fiche. A study was conducted by the staff of the Reconnaissance and Intelligence Data Handling Branch, Rome Air Development Center, to evaluate various microforms. The conclusion reached was that the 4" x 6" microfiche format was best for technical document storage and retrieval in the Federal Government.[8]

In view of the increasing use of microfiche by U. S. Government agencies, it is not surprising that during the past year the influx of documents to the AMIC system has changed from being almost 100 percent hard copy to about 65 percent now being received in the form of microfiche. This occurrence has had wide ranging effects. Changes have been necessitated in clerical handling and storage techniques, in the physical process of indexing, and in the reading of documents in microfiche form rather than as hard copy by the ultimate users of the system. Certain problems have arisen in dealing with microfiche, but certain advantages have become evident also. As experience with microfiche has increased, many of the problems have been resolved.

The University and the AFML have long recognized the efficacy of microform for certain AMIC operations. In 1966, the University purchased two microfiche readers and two microfilm reader-printers capable of handling both 16 mm and 35 mm film reels. As described in Reference 4, microfilm records were prepared of index and abstract cards for all the documents processed into the system. These records have proved valuable in screening search outputs and in providing a permanent reference record of bibliographic information, abstracts, and indexing for every document in the system. The reader-printers have been used for viewing documents which are available in reel microfilm form and for preparing hard copy of microfilmed abstracts.

In 1966 the AFML purchased a Recordak Miracode system, a self-contained microfilm document retrieval system. The AMIC system, which is maintained on an IBM 7094 at WPAFB, is essentially duplicated in the Miracode system. Although the Miracode does not have the searching capability of the computer, the index and abstract cards can be retrieved immediately, thus permitting on-site screening. The Miracode has been used when an immediate response to a technical request is required. In addition to the AMIC document file, the Miracode system is used for files of the Commerce Business Daily and Independent Research and Development reports. These items are indexed with AMIC vocabulary terms. The Atomic Energy Commission (AEC) operates a Miracode system for its files on physical properties of pure materials. The complete AEC files were obtained for the AFML and are available for searching in response to technical requests.

To supplement selective dissemination of information (SDI) services provided to AFML personnel, a subscription to Pandex was made. Pandex covers 2000 major journals in all areas of pure and applied science and provides full titles, bibliographic references and authors. The Pandex is provided quarterly in microfiche form.

This background and familiarity with microform was useful in dealing with documents in microfiche form when they started being received in quantity. Microfiche documents are now being received from NASA, AEC, and DDC on automatic distribution according to subject profiles registered with those organizations. In addition to automatic distribution, specific documents for inclusion in the AMIC system are ordered from these and other organizations in microfiche form whenever possible. Many previous documents stored in the AFML library as hard copy are being replaced with microfiche.

To process microfiche documents, it was necessary to obtain microfiche reading equipment. It was desired to evaluate various equipment specifically with regard to its usability by indexers and for occasional reading use. It was also anticipated that microfiche readers would have to be provided at a number of locations throughout the AFML for the use of its personnel. It was desirable to have equipment which would provide excellent readability and would be easy to use. However, the requirements for a suitable microfiche reader for indexers were considerably more stringent. Indexing from microfiche requires the indexer to sit in front of the reader for relatively long periods of time. He must change microfiche frequently. The reader should be compact and should fit easily on the indexer's desk. The indexer must not only scan and read the microfiche images but he must record index terms on an index card simultaneously. Furthermore, he must have ready access to the thesaurus of terminology. Portability was also a desired feature, since some part-time indexers index at home as well as in the office.

With the potential purchase of a sizeable number of microfiche readers, it was decided to purchase readers from several manufacturers and to evaluate each of them under actual use conditions. Five different microfiche readers were purchased and evaluated by several indexers. Each indexer participating in the evaluation was instructed to spend about an equal amount of time with each reader and to index for at least one period of four hours at a time with each one. The equipment was evaluated with regard to the following parameters: resolution, noise, tracking, viewing ease, focusing, construction and portability. A summary of the comments of the evaluators is presented in Table II (pages 8 and 9).

In general the optical systems of all readers seemed to be quite acceptable. The viewing screens of most of the readers were such that ambient light was reflected and was bothersome to the viewer The mechanical aspects of all the viewers also left quite a bit to be desired. For two readers in particular, the slightest contact caused them to go out of focus. The tracking mechanism which controlled image to image movement ranges from very poor to very good. The ruggedness of construction was variable, but most of the readers were soundly constructed. For indexing purposes, the readers with angled viewing screens were much easier to use since the indexer could record terms with a minimal alteration of body position. Changing of microfiche was relatively easy with all of the readers. One reader had outstanding portability. The other readers could be moved over short distances, but were unsuitable for transporting from office to home without risk of damage.

At the end of the evaluation, it was determined that Reader C was the best for our purposes. This reader was also considered very good for occasional reading use. To date, twenty-four of these readers have been purchased of which five are at the University and nineteen have been distributed to various locations at the AFML.

The microfiche form has the obvious advantage of compactness for storage, but certain problems in the processing and indexing of microfiche must be considered. Care must be exercised to assure that the microfiche itself does not become separated from the jacket which is marked with the access number. The compact size of the microfiche requires special care in handling to prevent misplacing or losing.

A typed abstract is required for every document entered into the system. Several techniques by which hard copy abstracts from microfiche could be prepared were tried. These included: typing directly from the microfiche using one of the portable microfiche readers, typing from hard copy abstracts found in announcement media such as Scientific and Technical Aerospace Reports (STAR) and U. S. Government Research and Development Reports

(USGRDR), and printing hard copy abstracts from the microfiche using the Recordak Magnaprint reader-printer with subsequent typing. Although the Magnaprint is designed for reel film, it is possible to maneuver the appropriate microfiche image into position for printing hard copy. Printing hard copy from the microfiche was found to be the most efficient of the three methods.

Indexing from microfiche is now being performed on a regular basis using the portable microfiche readers. Comments were solicited from experienced indexers regarding indexing from microfiche as compared and contrasted with indexing from hard copy. Of the five indexers surveyed, two indicated a preference for hard copy, two preferred microfiche and one indicated no preference for one over the other. All indexers felt that reading from a microfiche reader caused more eye strain than reading from hard copy. The primary objection in this regard is reflections from ambient lighting in the reader screen. By judicious placement of the readers, these reflections can be minimized. Another objection is the necessity to maintain essentially the same position throughout the day while indexing. One indexer mentioned that lack of image rotation tended to cause one to skip over charts and tables which appear vertical to the normal reading position. The same indexer felt that looking at microfiche on the reader was boring. Another indexer suggested that indexing from microfiche on the reader was psychologically enhanced because the microfiche are easier to handle and give the illusion of being smaller than bulky hard copy. The two indexers who prefer microfiche stated that they can index faster from microfiche than hard copy. One of the indexers who preferred hard copy indicated that he could index faster from hard copy. He also felt that he tired much less from indexing hard copy. The other two indexers did not find any significant time difference in indexing between the two forms. From this study it can be concluded that the indexers reacted to the indexing from microfiche vs. indexing from hard copy in a personal way. There were actually two indexers who preferred indexing from microfiche. One of those who was opposed to indexing microfiche reacted against this procedure vigorously, almost with the fervor of a crusader. Indexing from microfiche tended to become a more acceptable procedure as the indexers became accustomed to working with the readers. Presumably better equipment would enhance the acceptability of microfiche, but even with present day equipment the microfiche form was found in general to be acceptable to indexing personnel.

The reaction of AFML personnel to microfiche has been generally favorable. These users are occasional readers of microfiche and are not as subject to fatigue and eyestrain as indexers. The providing of good quality equipment in numerous convenient locations has been instrumental in gaining acceptance of documents in microfiche form by the engineers and scientists of AFML.

TABLE II. EVALUATION OF MICROFICHE READERS

Parameters	Reader A	Reader B
1. Resolution	Good. Outside light does not interfere with readability	Fair
2. Noise	No noise	Blower causes noise at an objectionable level
3. Tracking	Good, but care must be taken to avoid drifting	Very poor. Mechanism consists of easily scratched plastic plate on felt runners, slips and falls out readily
4. Viewing Ease	Very good. Easily readable	Poor. Machine is so constructed that it is necessary to insert one's head in a most uncomfortable position into the machine to read the fiche
5. Focusing	Good. Focus is uniform and holds from frame to frame	Very poor. Lens is held by only one set screw which loosens on focusing. Slightest jar knocks it out of focus
6. Construction	Very Good. Glass plate on top of machine appears vulnerable to breakage. Body of machine very rigid	Very Flimsy. This machine is very cheaply made and would not hold up
7. Portability, Compactness	The most compact of all the readers. Could be used on typist's desk for typing directly from fiche. Glass plate vulnerability limits portability	Good. The reader folds up to a compact size, is light and readily portable

TABLE II. CONTINUED

Reader C	Reader D	Reader E
Very good. Some interference from outside light, but this can be minimized with careful placement	Good	Very good. Outside light does not interfere
No noise	No noise	No noise
Very good. Moving from frame to frame is easy and frame remains stable between adjustments	Good	Very good. Rows lock into position and frame to frame movement is easily controlled. Frame remains stable between adjustments
Very good. Angle of screen particularly good for simultaneous reading and indexing	Satisfactory. Screen is angled slightly	Good. Screen is nearly vertical which is somewhat inconvenient for reading; it is also unsatisfactory for indexing. for best readability the machine must be on a low table
Good. Focus is sometimes slightly different for different locations on the same frame. Focus sometimes changes slightly from frame to frame, but focus adjust is easy	Fair. Easily knocked out of focus by slight jar. Focus can change from frame to frame	Very good. Focus is uniform and holds well from frome to frame
Excellent. This machine is soundly constructed, sturdy in use and is esthetically pleasing	Good.	Excellent. This machine is very sturdy. It is heavy enough to stay in place when in use
Excellent. The reader folds up into a vinyl covered case and is readily portable with out danger of damaging the machine even for house to office trips	Good. This unit is compact and light weight	This reader is not portable. It can be moved by one person for very short distances

Section IV

OPERATION OF THE AMIC DOCUMENT RETRIEVAL SYSTEM

Input. During the past year an increasing number of documents have been acquired for inclusion in the AMIC system by automatic distribution from the AEC, NASA and DDC. These documents are received in accordance with subject profiles registered with the distributing agencies. The documents are screened before they are forwarded for indexing and inclusion in the system. The effectiveness of the subject profiles for distribution is indicated in Table III, which shows the percentage of documents actually entered into the system from the total received from each agency. In addition to automatic distribution from AEC, DDC, and NASA, reports are received from in-house research, from AFML contractors, and by ordering documents dealing with materials research from other announcement media.

A number of handbooks which deal with some aspect of materials are received by the AFML. Handbooks contain useful information, and it was desired to put handbooks into retrievable form. To index handbooks in the same degree of detail as is performed for regular research documents was considered to be inappropriate. Such detailed indexing would require excessive indexing time and would result in many false retrievals. The establishment of a separate Miracode file for handbooks which would be indexed with general terminology was considered. Handbooks were assigned special access numbers designated by an "H" preceding the numerical element. Indexing was performed using general terminology e. g., for a handbook on aluminum alloys, the term ALUMINUM ALLOYS would be used for indexing rather than a detailed listing of all aluminum alloys included in the handbook. Certain terms which were especially useful for handbooks were added specifically to accommodate their indexing. To facilitate inclusion of handbooks in to the system, it was decided to enter the handbook indexing directly into the computer system rather than the Miracode which would be difficult to update. The retrieval of handbooks on search output is readily recognized by the special "H" designation of access number.

During the period covered by this report, 1 December 1967 through 30 November 1968, approximately 6000 documents were indexed and processed into the system. The documents were indexed with an average of 182 terms per document (exclusive of automatic generic postings) with an average indexing time of 30.7 minutes. Distribution by subject category is shown in Appendix I. There are now approximately 37,000 documents in the AMIC document retrieval system.

TABLE III. DOCUMENTS ACTUALLY ENTERED INTO AMIC SYSTEM FROM THE TOTAL RECEIVED ON AUTOMATIC DISTRIBUTION

Agency	No. Doc.'s Rec'd (Excluding Bibliographies, Classified, Foreign, Information)	% Entered into AMIC System
AEC (Since 5 Sep 1967)	3496	88.0%
NASA (Since 19 Oct 1967	1590	70.7%
DDC (Since 27 Feb 1968	1927	83.0%

Searching. A total of 300 technical requests were processed by the Information Systems Section during the report period. This represents an increase of 5% over the previous reporting period. An average of 8.7 abstracts was printed per search from the microfilm records and these abstract copies were forwarded with the search results to the technical requestors.

Figure 1 presents the total number of requests processed by the AMIC document retrieval system since 1963. The number of search requests performed for the AFML and for all other organizations is indicated in addition to the search totals. The number of search requests by the AFML and requests from outside organizations both increased slightly.

A listing of search requests by subject category is presented in conjunction with a listing of documents input to the system by subject category in Table V of Appendix I. A comparison of the two lists indicates that subjects of the documents entered into the system correlate reasonably well with the subjects requested. The relatively large number of physics documents reflects documents received on automatic distribution from the AEC.

One philosophical question regarding search requests which had never been resolved was the frequency with which the original requestors should be contacted by the information specialist who frames the search strategy to retrieve the documents pertinent to the request. One thought is that all technical requests should be checked with the original requestor regardless of the apparent clarity of the request. At the opposite end of the spectrum are those who feel that the information specialist who knows the system is best qualified to interpret what the requestor really means and should proceed without referring back to the original requestor. At the University, an intermediate approach was taken. If the request appeared to be straight forward and clear, the information specialist set up the search strategy without checking with the originator. For example, a request on the thermal conductivity of S-glass reinforced Epon 828 at 350-500°F would be considered a request which would not require further checking. If the request as stated were not clear, if it would result in a large number of retrievals, or if it were subject to variations in interpretation, the original requestor would be contacted if possible. A request was received in which mechanical properties of composites was desired. Such a request taken at face value would result in numerous retrievals. Checking with the original requestor, it was found that he was interested in nonablative structural polymer matrix composites with any type of reinforcement. Fatigue life, creep and flexural strength and modulus of elasticity were the properties of particular interest. He was interested only in information not more than five years old. This additional information was necessary to formulate a reasonable search strategy.

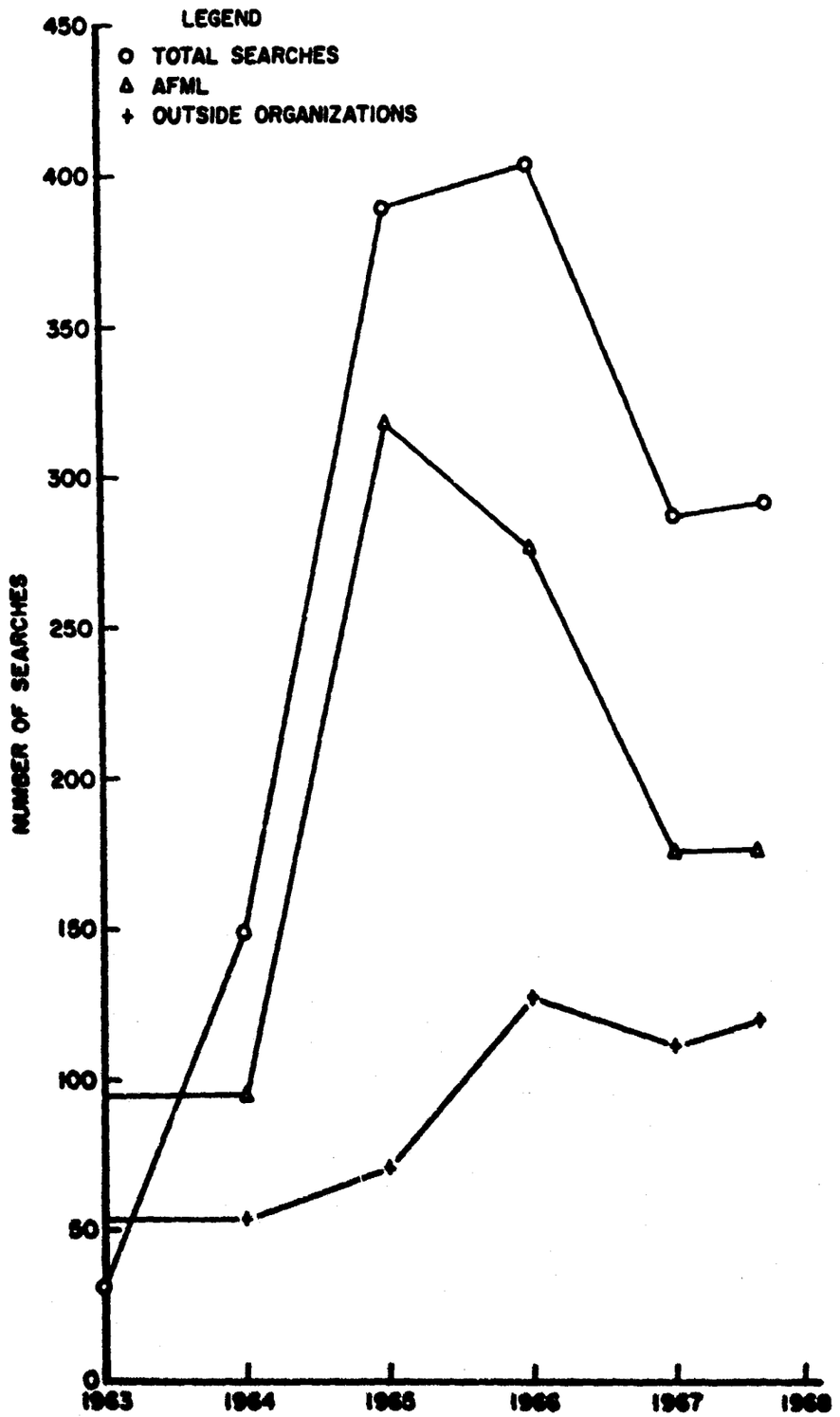


FIGURE 1: SEARCH REQUEST PROCESSED BY YEARS 1963 - 1968

To determine the desirability of always checking with the original requestor, a study was performed in which ten consecutive questions were referred back to the requestors regardless of the apparent clarity of the requests. Search strategies were formulated before checking with the requestors to see whether the search strategies would require modification. Of the ten requests, only one normally would have been checked back for additional information. For this one situation, additional information on the temperature range of interest aided the information specialist in eliminating many false retrievals which otherwise would have been obtained. Of the remaining requests, two search strategies were modified and the other seven remained unchanged.

The reactions of the original requestors to being called were interesting. Six of the requestors were AFML personnel and four were from outside organizations. The person who would have been contacted was cooperative in clarifying his request. One individual was rather irritated at being called. He explained that he had already gone through a discussion of his request with an AFML information specialist and had thought that a clear statement had been made. That request would not normally have been referred back. The procedure of contacting the original requestor caused a delay in processing the search request. The requestor was often not in and messages left to call the University were not acted upon immediately. Sometimes it was necessary to call him again.

The results of this study substantiated the University's practice of contacting the original requestor only when judged by the information specialist to be necessary. In this way most requests can be processed immediately. If it should be determined that the search results were not appropriate to the request due to misinterpretation, the requestor could have the search rerun with a different search strategy. This has been done in several instances. Advantages of this procedure are that the request is processed immediately and the requestor can see the first search strategy and the documents retrieved. With this information, the requestor can interact more effectively with the information specialist in setting up the search to be rerun. Furthermore, at least some pertinent retrievals might be provided right away. A disadvantage is that searches are expensive to run, and prior clarification might prevent an otherwise abortive search. It should be noted that the requestor might lose confidence in the document retrieval system and the personnel responsible for its operation if he is called back to clarify or verify every request. All factors considered, the practice of clarifying requests only when judged necessary by the information specialist appears to represent the best balance between delaying the request processing and the risk of running an inappropriate search.

Thesaurus. The thesaurus was updated once during the reporting period. The collection terms introduced in the previous updating proved to be useful, and a number of additional collection terms were included in the most recent updating. A collection term is one in which a group of terms related to a common concept or having a common element are displayed in one location. For example, the term "energy" does not exist as an active term, but there are a number of active terms such as ELECTRICAL ENERGY, KINETIC ENERGY, etc. The collection term ENERGY shows all the various forms of energy which are active terms. A number of vocabulary and thesaurus additions were made to the main system. In particular, additional composite terminology was added based on experience with indexing of documents on composite materials and search requests in which information on composite materials was desired.

An introduction section was made an integral part of the thesaurus. This introduction indicates certain standard procedures for listing items for the AMIC system and gives an explanation of active terms, reference terms and collection terms. The introduction is designed to aid indexer trainees and others unfamiliar with the AMIC system.

Consideration was given to ways in which the thesaurus could be made easier to use. One possibility was the separation of the thesaurus into more than one alphabetically ordered section, each section representing a logical grouping of terminology. This type of display would present all terms belonging to a group in close proximity rather than interspersed with terms of other areas. The problems created by separating the straight alphabetic listing into sections had to be considered as well as the advantages. In the first place, any division must be somewhat arbitrary. One could divide according to discipline, e.g., physics, chemistry, mathematics, etc.; by engineering considerations such as materials, processes, properties; or by some other scheme. Whatever division technique is selected, the divisions or categories must be mutually exclusive to be effective. If they are not, one term could conceivably fit into more than one category or microthesaurus, and the user would be forced to look in several places. This procedure would be more burdensome than looking through a single alphabetic listing. The more separation which is performed, the more care must be taken to assure mutual exclusiveness of the categories. For example, if all metallurgical related terminology were to be separated out, the term HEAT TREATMENT would require consideration. If HEAT TREATMENT were considered as belonging to the metallurgical category, how would one handle HEAT TREATMENT of a synthetic polymeric fiber?

Careful consideration was directed to the desirability of performing separation of the AMIC thesaurus. Since metallic materials terminology constitutes approximately one-third of all terminology, it was decided that it might be useful to separate the AMIC thesaurus into two sections - one on

metallic materials and the other containing all other terminology. This division was designed to make it easier to locate terms and to group all metallic materials terminology. For example, with straight alphabetic listing, it is necessary to page through a long list of magnesium alloys to get to magnetic properties terminology.

To test the usefulness of dividing the present AMIC thesaurus into metallic materials and all other terminology, the most recent update provided some thesauri divided into the two sections and some thesauri in the normal straight alphabetic listing. Certain indexers were provided with both types of thesauri and instructed to use each of them about equally. The reactions of the indexers were favorable to the separated thesaurus. One indexer commented that it took some time to get used to having the metallic materials terminology located in the back section of the thesaurus, but after he remembered its location, the time required for looking up terms was less using the separated thesaurus. Other indexers either preferred the separated format or indicated no preference.

Chemical Abstracts Service. The University has continued subscriptions to CAS SDI programs to provide improved information services to personnel in the AFML. The services subscribed to are Basic Journal Abstracts (3JA), Polymer Science and Technology from journals (POST-J), Polymer Science and Technology from patents (POST-P) and CA Condensates. These SDI services are particularly of value to the AFML Polymer Branch (MANP). The first three services involve searching the titles and texts of abstracts for technical terms contained therein according to prescribed user profiles. The fourth service, CA Condensates includes titles, authors' names, journal and/or patent references, and CAS issue Keyword Index entries for each abstract in the published versions of Chemical Abstracts. An estimated 240,000 abstracts will be published in CAS in 1968. CA Condensates involves searching by the Keyword Index entries rather than searching titles and abstract texts; however, that searching has just been started.

Information Science Documents. A number of information science documents are on hand at the University and new accessions are added continually. These documents constitute the basis for establishing a document retrieval system for information science documents. The documents of primary interest to the AFML deal with information storage and retrieval, but the area of information science is much broader in scope. To restrict the indexing of documents only to those in the area of storage and retrieval would eliminate many documents in other areas which could be of interest. On the other hand, indexing all areas in detail would present problems of vocabulary control, terminology definition and inadequacy in some areas. A decision was made to index in detail the documents on information storage and retrieval and to establish general categories with perhaps a limited number of vocabulary terms applicable to areas of information science other

than information storage and retrieval. A review was made of various sources covering the broad area of information science. Catalogs from graduate schools which offer programs in information science were particularly helpful in providing a range of subject matter which constitutes information science. Main topics tentatively selected are as follows: Application, Automata Theory , Communication Theory, Computer Systems (Software), Decision Theory, Education, Engineering, Information Storage and Retrieval, Information Systems, Information Theory, Linguistics, Logic Design, Man-Machine Interface, Mathematics, Physical Devices (includes computers), Simulation (Artificial Intelligence) and Social Aspects.

Free indexing has been employed for vocabulary generation for the area of information science. Documents are indexed by two individuals, one having a background in information storage and retrieval and the other being a graduate student in the Information Science Department at the University of Dayton. The indexing of the documents is compared and differences are resolved. Agreement has been achieved on a selected group of documents as to the concepts to be indexed, and standardization of terminology to express the concepts is underway. Preliminary thesaurus building efforts in which term relationships have been considered have taken place. Although no viable thesaurus has been developed as yet, these efforts have been most useful in establishing meaningful concepts and terms.

The information science documents are maintained in a separate storage location at the University. Access numbers have been assigned to all of the documents on hand. Cross reference card files have been established which identify documents by author, AD number, and access number.

Personnel Time Distribution. The time spent by personnel on this contract has been broken down into various task numbers. The task number definitions are shown in Figure 2 and the personnel time distribution for various types of personnel are indicated in Table IV.

FIGURE 2

DEFINITION OF TASK NUMBERS

Task No.	Function	Description
01	General	Includes:
		Supervision
		Meetings & trips
		Holidays & sick leave
		Writing of reports
		Training of Students
		Time spent with visitors
02	Input	Includes:
		Assignment of accession numbers
		Document accounting records
		Preparation of index and abstract cards
		Indexing
		Keypunching
03	Output	Includes:
		Preparation of search strategy
		Search
		Screening of searches
		Search accounting records
		Library loan functions
		Keypunching
04	Updating	Includes:
		Review of vocabulary and thesaurus
		Changes or additions to previous records
		Keypunching
		Acquisition of missing documents

FIGURE 2 (Continued)

Task No.	Function	Description
05	(UD) Research	Includes: Evaluation studies Studies of new techniques Investigation of new systems
05	(AFML) Library	Includes: Preparation of Materials Information Bulletin
06	Special Projects	Includes: Work performed in support of the AFML not directly related to AMIC retrieval system
07	Microfilming	Includes: Time spent on the microfilming of index/abstract records

TABLE IV. DISTRIBUTION OF PERSONNEL TIME BY TASK NUMBER

Professional and Clerical at UD:

Task Number	Percent of Time
01	22.7
02	42.3
03	6.8
04	3.6
05	8.5
06	15.9
07	0.3

Clerical at the AF Materials Laboratory Library

Task Number	Percent of Time
01	9.4
02	57.1
03	6.3
04	6.7
05	11.8
06	8.7

Section V

COMPUTER AIDED INDEXING

The University of Dayton has to date employed indexer personnel for the indexing of technical documents. There are organizations which have employed varying techniques of automated indexing. There are certain advantages to using indexers. Indexers are able to make discriminations as to the relative importance of technical concepts as they appear in an abstract or document. They have access to the entire document and they can go beyond the document itself to reference books, to consultation with experts, or other sources as deemed appropriate to aid in properly indexing the document at hand. Indexers can apply inductive reasoning to formulate and index concepts which are implied by the document but not expressly stated. Indexers become familiar with the requirements of the users of the system by participating in search request analysis, search strategy formulation and search screening. This user-information specialist interaction guides the information specialist in modifying indexing terminology and techniques to satisfy best the requirements of the users.

There are also disadvantages to using indexers. Because of the variations in backgrounds and personalities of people, different persons approach the indexing of a given document in different ways. Even the same indexer can approach the same document differently on different days. Despite these factors, excellent indexing results regarding indexer consistency have been obtained. The University used a number of part-time indexers including undergraduate students, graduate students and high school teachers. There is, however, a considerable turnover in personnel which entails constant training of new personnel. There is also a degree of repetitiveness in indexing which causes indexer fatigue, especially for long periods of indexing. By using part-time personnel, this problem is not as acute as would occur for full-time personnel, but even the part-time indexers are affected by the tediousness of indexing.

The computer aided indexing concept arose out of a desire to use the computer to perform the routine aspects of the indexing process while permitting the indexers to use their intellects to resolve non-routine situations. By allocating the treatment of commonly occurring and reoccurring indexing situations to the computer, the indexer is freed from making and remaking the same decisions when encountering them from one document to the next. The human function could then be upgraded to making decisions regarding only new indexing situations. As such decisions would be made, they could be added to the computer's capability. This higher level human activity would be more stimulating and thus would alleviate much of the tediousness

of the indexing function. On the other hand, a greater responsibility is placed upon the human decision maker, since the decision made would affect not only the indexing of the document at hand but also the indexing of all subsequent documents in which that indexing situation occurred.

Computer analysis of text is the fundamental step in computer aided indexing. In applying this step, text is entered onto core from a remote terminal or by batch processing and the text is compared with a file of nontechnical words such as THE, BUT, AND, etc., which convey no technical concepts. Words from the text which match nontechnical words are then disregarded for the remainder of the text analysis.

The text is next compared with a technical word file. If the word under consideration matches a word in the technical file, instructions are provided for storing the word under an appropriate "conceptual location" for subsequent retrieval. It should be pointed out that there could be a number of technical words which express a given indexing concept. For instance, the terms cloth, textiles, fabric, woven structure, felts, webbing and netting might all be considered as belonging to a particular concept. This concept might be called "fabric structures", but it would not necessarily require an English name. The concept could have a particular code number which could serve not only to represent that concept but to show its hierarchical and other relationships to other concepts.

Words which were not recognized by the computer as being either technical words or nontechnical words would be checked for possible bound term meaning. If this test fails they would be stored and printed out for human intellectual decisions. In contrast to the human indexing of documents in which the same decisions must be made repeatedly, once a decision had been made, it would become part of the computer aided indexing system. Any time the technical word was encountered in the future, it would automatically be treated according to that one-time decision. This feature overcomes one of the problems of indexing inconsistencies.

As mentioned previously, provision has been made for the situation in which two words are required to express a single concept, i.e., a bound term. A technical concept can be expressed by a combination of a technical word and a nontechnical word, by two normally nontechnical words or by two technical words which in combination express a different concept than would be expressed singularly. It is also possible that more than two words in combination could express a technical concept. The programs prepared allow only for the contingency of two word bound terms, but the capacity for handling more than two word bound terms could be built in. A file of bound terms and a file of words which may be parts of bound terms is necessary in conjunction with the technical word file and the nontechnical word file. An example of a bound term is the term TENSILE STRENGTH which expresses a single concept.

The word LIGHT is an example of a word which may be technical, nontechnical or part of a bound term. LIGHT by itself may refer to electromagnetic radiation in the visible region of the spectrum, may refer to daylight, may refer to artificial illumination or may be an adjective referring to the relative weight or density of an object. The bound term LIGHT METALS represents a concept. The expression ULTRAVIOLET LIGHT is a bound term referring to electromagnetic radiation in the ultraviolet rather than the visible region of the spectrum. The word THERMAL by itself is essentially meaningless, but as part of bound terms it is instrumental in conveying the concept, e.g., THERMAL GRADIATION, THERMAL INSULATION, THERMAL CYCLING, THERMAL EXPANSION, etc. In the text analysis, if a word matches one in the file of potential bound term elements, a program must check the word preceding and the word following the word under consideration in conjunction with the word under consideration to determine if the word being examined is part of a bound term. This is accomplished by comparing the resultant potential bound term with the file of actual bound terms.

The abstract file is the file which contains the text of the abstract in conjunction with the associated bibliographic information. The bibliographic information includes the internal access number, the number assigned by the issuing agency, such as the AD number, N number or other report numbers, title, author, contract number, contractor, sponsoring agency, project monitor, and the date of the report. In addition to this information each abstract has a key number which serves as an identification number from the computer standpoint.

This computer aided indexing system requires at least five files: the abstract file, a file of words which could be bound term elements, the bound term file, the nontechnical word file, and the technical word file. To summarize briefly the computer aided indexing concept, abstract text or other technical text is entered into the system and the text is analyzed for nontechnical words, technical words, and bound terms. Nontechnical words are disregarded. Technical words and bound terms representing technical concepts are converted to appropriate code numbers or AMIC terms which then represent access or retrieval parameters for the document. In addition to these technical retrieval parameters, retrieval can be effected by any one or more of the bibliographic parameters.

Although the computer aided indexing project was initiated to assist in processing documents into the AMIC document retrieval system, it was logical to consider establishing a complete self-contained document indexing, storage and retrieval package. The processing of technical requests could be performed by the same basic text analysis technique as used for document input processing. Requests could be presented in normal English text and the request text could be analyzed in much the same way as the abstract text

with a program to retrieve and display the relevant document citations, titles, abstracts, or abstract portions as desired. The computer aided indexing project was predicated on the availability of remote access capability for document input and retrieval. It is expected that the search requestors would interact directly with the computer system through a remote terminal. Programs would be prepared to establish conversational mode between the computer and the requestor to assist the requestor in clarifying and establishing limits on the scope of his request as he himself deems appropriate.

Although computer aided indexing has been rather well developed conceptually, the implementation of the concept to an actual operating system is a slow process. The preliminary work had been done on an IBM 360 Model 40 computer to which the University has had access. A 2741 typewriter communications terminal was used in conjunction with the 360. Communication was established between the 2741 terminal and the 360 central processing unit (CPU) using the basic telecommunication access method (BTAM) software package. The communication capability permitted the loading of program modules, the execution of programs and the input of data from the communications terminal. Output from the CPU could be directed either back to the terminal or to the line printer of the computer system.

File structuring was an important consideration in this effort because of the magnitude of the files. Text analysis proceeds word by word, and every word encountered must be compared with words contained in the files to locate a match if it exists. A random listing of words in the files could require a search of nearly all the files to locate the word matching the one under consideration. This procedure would require excessive computer time. Alphabetizing all the words would provide some structuring and would improve the search and match procedure, but it would still be too time-consuming.

To converge as rapidly as possible on a word match, a file structuring technique based on the length of the word and the alphabetization of the first two characters was employed.[9] Four direct access keyed files were structured in this way. To demonstrate the working of the file, let it be assumed that the word under consideration is FATIGUE. The word is recognized as a seven letter word the first two characters of which are FA. This analysis directs the word being considered to the appropriate file location extremely rapidly. A search of this small portion of the file, i.e., all seven letter words beginning with FA is sufficient to determine that either a match occurs or does not occur. This structure provides for locating a program manageable data string from files containing a great number of English words. This file structure was established for the file of technical words (TECH), for the file of nontechnical words (NONTECH), for the files of words which could be elements of bound terms (SWAMP) and for actual bound terms (HASH).

In addition to these four file structures, the structuring and construction of an abstract file (ABSTR) was accomplished. A limited number of abstracts including both abstract text and corresponding bibliographic information were entered from the terminal into the abstract file. Programs were prepared to enter the abstracts, to perform text searching, and to edit or update the abstract by adding, deleting or changing any portion of the abstract text or bibliographic information. Additional details and program descriptions are given in Appendix II.

Subsequent to the work just described, the University has acquired an RCA Spectra 70/46 computer the operation of which will be entirely under the direct jurisdiction of the University. This computer is scheduled for delivery in December 1968. Any further work with the computer aided indexing project will be performed using the RCA equipment.

Section VI

SUMMARY

Additional part-time indexers were employed and trained in accordance with the training program developed and in use at the University for about two years. The indexers were evaluated by the indexer performance analysis technique described in the previous annual summary report.[5] Results indicate that the percentage of essential terms and total terms relative to the experienced indexer's standard increased slightly over indexer trainees of last year. Particularly interesting was a significant decrease in the use of misleading terms, presumably due in large part to thesaurus improvements.

Documents are now being received in microfiche form, especially documents received by automatic distribution. The Bell & Howell Mascot microfiche reader provided a majority of desirable features for use by indexers and for occasional reading use. Several of these readers were purchased for use by indexers in indexing microfiche. Certain problems in dealing with microfiche were encountered. To obtain hard copy abstracts it was found that the microfilm reader-printers could be used to print hard copy from which the typist could prepare typed abstracts. This method was easier than typing from an abstract image on a microfiche reader screen or locating abstracts in announcement media such as TAB and STAR. Comments were solicited from indexers regarding indexing from microfiche. About one-third preferred hard copy, one-third preferred microfiche and the other one-third indicated no preference for one form over the other.

Input to the document retrieval system increased over the preceding year. More than 6000 documents were added to the system which now contains a total of 37,000 documents. Three-hundred technical requests were processed in the past year, an increase of about 5%. The number of technical requests from both the AFML and from outside organizations increased somewhat. The thesaurus was updated with new active terms, reference terms and collection terms. A division of the thesaurus was made into two parts: metallic materials and all other terminology. This separation has been beneficial in making the thesaurus easier to use. The CAS SDI services have been continued. A system for indexing, storage and retrieval of documents dealing with information science is under development. The scope of topics has been tentatively described. Main emphasis is being placed on information storage and retrieval.

A computer aided indexing scheme was devised to take advantage of the capability of the computer in performing the repetitive aspects of indexing. Human intellectual capability would be applied to resolving non-routine

situations, and the resultant decisions would then become incorporated as part of the computer's repertory. The basic concept consists of technical text analysis in relation to technical terms, nontechnical terms and bound terms which convey technical concepts. The technical terms and bound terms would be converted to technical concepts such as are used in the AMIC system or to code numbers which represent technical concepts. A logical extension of the concept of computer aided indexing scheme to a self-contained system including the functions of indexing, storage, and retrieval was made. Provision was made in the concept for direct interaction of users through remote access terminals with the computer system. Accomplishments in the actual implementation of the concept included the establishment of communication capability between the communications terminal and the CPU with BTAM software, the structuring of keyed direct access files for large numbers of English words and the creation of an actual abstract file which allowed text searching and editing/updating functions to be performed.

A recent development has been the acquisition by the University of an RCA Spectra 70/46 computer system scheduled for delivery in December 1968. This computer will be owned and administered by the University. Additional work on the running of CAS searches and the furthering of the computer aided indexing project will be accomplished using the RCA equipment.

REFERENCES

1. E. A. Janning, Establishment of a Coordinate Indexing Retrieval System for the Air Force Materials Laboratory, RTE-TDR-63-4263 Air Force Materials Laboratory, N-RAFB, Ohio, November 1963. (AD 428 423)

2. E. A. Janning, The Modification of an Information Retrieval System by Improving Vocabulary Control, Indexing Consistency, and Search Capabilities, AFML-TR-65-20, Air Force Materials Laboratory W-PAFB, Ohio, March 1965. (AD 613 301)

3. E. A. Janning, Operations of a Document Retrieval System Using a Controlled Vocabulary, AFML-TR-66-36, Air Force Materials Laboratory W-PAFB, March 1966. (AD 633 614)

4. F. L. Scheffler, Student Indexer Training Program and the Improved Operation of a Document Retrieval System, AFML-TR-66-391, Air Force Materials Laboratory, W-PAFB, Ohio, January 1967. (AD 651 039)

5. F. L. Scheffler, Indexer Performance Analysis and Operations of a Document Retrieval System, AFML-TR-67-379, Air Force Materials Laboratory, W-PAFB, Ohio, February 1968. (AD 666 462)

6. A. Teplitz, Library Fiche: An introduction and Explanation, SP-29922/000/01, Systems Development Corporation, Santa Monica, California, October 1967. (AD 661 660)

7. J. T. Salisbury, A Study on the Application of Microfilming to the Production, Distribution, Use and Retrieval of Technical Reports, 65SD249, General Electrical Company, Philadelphia, Pennsylvania, April 1965. (AD 615 800)

8. Anonymous, A summary of the State-of-the-Art in Microfilm Document Storage and Retrieval Systems, RADC-TR-67-496, Rome Air Development Center, Griffiss AFB, New York, September 1967. (AD 820 127)

9. T. Angel and J. Anderson, Seminar on File Structures for On-Line Systems, Association for Computing Machinery, Sheraton Hotel, Columbus, Ohio, 4 March 1968

APPENDIX I

SUBJECT CATEGORIES

AMIC	COSATI	CATEGORY
01	01	Aeronautics
		Aerodynamics
		Aeronautics
		Aircraft
		Aircraft flight control and instrumentation
		Jet engines
02	03+04	Astronomy, Astrophysics, Atmospheric Sciences
		Astronomy
		Astrophysics
		Atmospheric physics
		Meteorology
03	06+07	Chemistry, Biology, Medical Sciences
		Biochemistry
		Bioengineering
		Biology
		Chemical analysis
		Chemical engineering
		Inorganic chemistry
		Life support systems
		Organic chemistry
		Physical chemistry
		Radiochemistry
		Toxicology

AMIC	COSATI	CATEGORY
04	09	Electronics and Electrical Engineering
		Components
		Electronic & electrical engineering
		Telemetry
05	11A	Adhesives
		Ceramic cements
		Organic resin adhesives
		Potting compounds
06	11A	Seals, Sealants
		Ceramic-metal bonds
		Mechanical seals
		O-rings
07	11B	Ceramics, Refractories, Glasses, Minerals
		Borides
		Carbides
		Carbon, graphites
		Mixed oxides
		Nitrides
		Single oxides
08	11C	Coating, Paints, Oxide Films
09	11D	Composites Materials, Laminates, Sandwich Structures, Honeycomb
10	11E	Fibers, Textiles, Cloth
11	11F	Metallurgy, Metallography
		Alloys
		Metals
12	11H	Oils, Lubricants, Heat Transfer Fluids, Greases, Hydraulic Fluids
13	11I	Polymers, Plastics

AMIC	COSATI	CATEGORY
14	11J	Elastomers
15	11K	Cleaning Compounds, Surface Active Agents
16	11L	Wood and Paper Products
17	21	Fuels, Propellants, Propulsion Systems, Explosives
18	13	Mechanical, Industrial, Civil and Marine Engineering
		Civil engineering
		Construction equipment, materials, supplies
		Containers and packaging
		Coupling, fittings, fasteners, joints
		Industrial processes
		Machining, tools, machine elements such as bearings, gas lubrication systems
		Marine engineering
		Pumps, filters, pipes, fittings, tubing, and valves
		Safety engineering
		Structural engineering
19	14	Methods and Equipment
		Apparatus
		Detectors
		Laboratories, test facilities, and test equipment
		Recording devices
20	18	Nuclear Science and Technology
		Fuel elements; fuel, nuclear
		Nuclear explosions
		Nuclear power plants
		Nuclear reactors

AMIC	COSATI	CATEGORY
20 (Con't)	18	Radiation shielding
		Radioactive wastes
21	20	Physics
		Acoustics
		Crystallography
		Electricity and magnetism
		Fluid mechanics
		Masers and lasers
		Optics
		Particle accelerators
		Particle physics
		Plasma physics
		Quantum theory
		Solid mechanics
		Solid-state physics
		Spectrometry, spectroscopy
		Thermodynamics
		Wave propagation
22	10, 16, 22	Space Technology and missiles
		Astronautics
		Energy conversion, solar cells
		Launch vehicles
		Missile technology
		Re-entry vehicles
		Rockets
		Satellites, artificial
		Spacecraft
		Trajectories and re-entry

TABLE V. DOCUMENT INPUT AND SEARCHES PROCESSED BY SUBJECT CATEGORY

AMIC CATEGORY	DOCUMENT CATEGORIES		SEARCH CATEGORIES	
	No.	%	No.	%
01	106	1.5	23	5.6
02	40	0.6	3	0.7
03	848	11.8	30	7.3
04	185	2.6	19	4.6
05	29	0.4	5	1.2
06	28	0.4	0	0.0
07	219	3.0	24	5.9
08	141	2.0	14	3.4
09	400	5.6	37	9.0
10	136	1.9	12	2.9
11	1103	15.3	106	25.9
12	319	4.4	3	0.7
13	378	5.2	30	7.3
14	61	0.8	8	2.0
15	13	0.2	0	0.0
16	15	0.2	1	0.2
17	97	1.3	2	0.5
18	519	7.2	46	11.2
19	370	5.1	11	2.7
20	246	3.4	3	0.7
21	1665	23.2	31	7.6
22	269	3.7	6	1.5

APPENDIX II

Program Descriptions and Hardware Configuration for the Computer Aided Indexing Project

A description of the computer aided indexing project concept was presented in Section V of the body of the report. In this appendix the structure of the computer aided indexing system will be presented. This structure consists of the description of the functions to be performed, the files used and/or created by that function, and the programs that accomplish that function.

To demonstrate the operation of the programs and files as they function to process an abstract into the system, a hypothetical access number 28,012 and a hypothetical term, BRAKING, will be considered. The abstract text of 28,012 contains the word braking. It is assumed that the term BRAKING has been assigned the ten character number 5700000002. The structure of the number reflects a similarity to the number 5700000003, but there is some distinguishing characteristic between the two, also. Note that the ten character numbers represent concepts rather than terms, e.g., more than one English word may have the same ten character number as long as the concept conveyed is the same. In the example chosen, BRAKED, TO BRAKE, etc., may also have the term number 5700000002.

It should be pointed out that there are three different key numbers which will be referred to throughout the remainder of this example. The term number or term key number has just been discussed. This number represents the technical concept. There is also a key number generated for the English word itself which depends on the word length and the first two alphabetic characters. This key number is called the word key number. The third key number corresponds to the abstract access number and serves to identify the abstract. This number is called the identification key number.

When the system recognizes the term BRAKING as it is typed into the computer, a key number is constructed which is based on the length of the word and its first two characters. The direct access file NONTECH is then scanned by looking at the terms listed under the key number corresponding to BRAKING. In the case under consideration, the term would not be found in NONTECH, since BRAKING conveys a technical concept. After it has been determined that the term BRAKING is not on NONTECH, the file TECH would be direct accessed by the same word key number. The key number would be found in TECH. Associated with the word key number (for BRAKING) is a ten character identification key corresponding to an access number. This is the key number of the first abstract in the file indexed by that concept.

Using this identification key number, the file ABSTECH, which contains a list of identification key numbers including the one which is being sought is accessed. Associated with this identification key number are term key numbers for each of the technical concepts that were indexed for that document. This list of term key numbers is scanned until the number 5700000002 is located. Associated with this term key number is the identification key number of the next abstract indexed by this concept. This identification key number is searched for in the ABSTECH file, and the term key number 5700000002 is found which in turn leads to the next abstract identification key number indexed by that concept. This procedure is repeated until the last document that was indexed by that concept is found. At this point the abstract identification key number corresponding to 28,012 in association with the term key number is entered. This procedure will lead the next search to the abstract which is being dealt with, namely, 28,012. The abstract text and bibliographic data is then entered under the appropriate identification key number in the file ABSTR. The abstract has thus been automatically indexed under the term BRAKING 5700000002 and the abstract for 28,012 has been entered into the ABSTR file. The abstract is entered into the ABSTR file immediately after the first technical word is encountered, in this case, the term BRAKING.

If the term under consideration should be one of the type which does not inherently convey a technical concept, i. e., it is not found on NONTECH or TECH, the SWAMP file must be checked to see if the word in combination with another word could convey a technical concept. If the word appears in SWAMP, the file of bound terms called HASH is checked. If the word is located in HASH, there is a term key number corresponding to the bound term, and the abstract is entered as described for a word appearing in TECH. The program permits combining the word preceding and the word following into a bound term concept. If the word does not appear in NONTECH, TECH, or SWAMP, it is presented to the human indexer for a decision. This decision must be reflected by updating the appropriate files.

It should be noted that the file structure indicated represents the simplest version of the file concepts. Considerable sophistication of file structure would be possible and would require study as the conceptual ideas moved into the implementation phase.

I. CREATION AND MAINTENANCE OF THE ABSTRACT FILE

PURPOSE: To list each abstract for easy accessibility at retrieval time.

FILES: ABSTR

ABSTR, a direct access file, contains each full text abstract, and each entry is assigned a unique key number. The length of the abstract is limited to 1630 character positions. In a 2314 disc configuration, four full text abstracts can be accomodated on each track.

PROGRAMS: SAB, RENEW

SAB creates and makes additions to ABSTR file. The abstracts can be read in on cards or typed in from a 2741 terminal. The abstracts are divided into eleven parts, each followed and separated from the others by an asterisk (*). The eleven parts of the abstract appear on the file in the following order: file key #, access #, title, report #, date, author, corporate author, sponsor, contract #, body, and terminating key #. Should any part be unavailable, its place is held in the file by its trailing asterisk. More parts, such as the AD number, could be added as desired.
RENEW updates and deletes abstracts already listed in ABSTR. Each abstract to be corrected is read into core, broken down into its parts, and then reassembled with the erroneous or absent part or parts replaced by the updated part or parts, which can be read from cards or from a terminal. The updated abstract is then written on the file. A whole abstract can be deleted by replacing each of its parts with null strings.

II. SEARCH AND RETRIEVAL TECHNIQUE

PURPOSE: To retrieve part or all of any abstract which may contain any information relating to the topic or topics designated by the user.

FILES: MASTER, ABSTECH

MASTER, a direct access file, holds a ten character position for every technical concept or term anticipated in an abstract. This position is held by a word code number which may, in the future of this system, relate the word which it represents to another word or group of words which together make up a technical concept. The code number position held in MASTER contains the identification key number of the first abstract listed in ABSTR in which the term appears. Many of the positions will be blank in anticipation of words to be encountered in future abstracts. The key number in this position initializes the process by which terms are indexed into the threaded file, ABSTECH.

ABSTECH, a direct access file, is not only a list of the term code numbers of all the corresponding technical terms found in an abstract, but is also a threaded file of indexed information. Each abstract in ABSTR is represented in ABSTECH by a string of term code numbers which make up the forementioned list. Each term code number is separated from the others in the list by a space and is followed by a ten-character field. This field is blank if the term corresponding to that particular code number has not been detected in any abstracts added to file ABSTR after the abstract in question had itself been added. Otherwise, this field contains the identification key number of the next abstract in ABSTR in which that term appears.

PROGRAMS: SEARCH

SEARCH handles the retrieval of part or all of any abstract in the ABSTR file. After reading, from cards or from a terminal, the abstract key number of the abstract to be retrieved and a list of parameters describing the parts of that abstract to be printed, SEARCH pulls the requested abstract off the disc and breaks it down into its eleven parts. The parameters are interpreted and the appropriate parts of that abstract are printed.

PROGRAMS: (Continued)

While the programming of the searching technique, which would normally precede the use of SEARCH, has not yet been developed, the philosophy of such a technique has been firmly imbedded in the file structure of MASTER and ABSTECH. After determining the topic of the search by reducing a user's search question to a few basic technical concepts, the concepts are transformed into term code numbers by files TECH and SWAMP (see part III). Each term code can be located in MASTER, where one finds the identification key number of the first abstract in which that term code's corresponding term appears. Upon moving to that abstract's ABSTECH list, the next abstract in which the term appears can be found by locating that term's code number in the list and pulling off the abstract key number which follows the code number. A search through this second abstract's ABSTECH list reveals the third abstract in which the term appears, and so on through the threaded file. A list of the identification key numbers of all the abstracts in which this term's code number was located may be assembled for each term or concept involved in the search topic. Full logical capabilities may be applied to these lists of abstract key numbers to reduce the number of abstracts to a workable amount, at which time SEARCH could be employed.

III. INDEXING TECHNIQUE/MAINTENANCE OF VOCULARY AND THESAURUS FILES

PURPOSE: To provide both a base and a technique for the automatic indexing of abstracts.

FILES: NONTECH, TECH, SWAMP, HASH

NONTECH, a direct access file, contains every word which is not normally attributed technical significance. Each word is grouped with certain other words according to its length and the alphabetic sequence of its initial two characters. In this setup, a small string of words with identical length and initial two characters can be directly accessed and searched, thus eliminating the problem of searching the entire file for a term. All words of more than ten characters in length and having identical initial two characters are grouped together as one string, as are all non-technical words which begin with numerals. Each string is 550 characters in length with entries in each string being separated by a space. Ten strings reside on one 2314 track. TECH contains all words which, standing alone, usually convey technical ideas and associated with each word term code number, which relates each word to a technical concept. The grouping of words is similar to the structuring of NONTECH, the difference being the separation of words up to 15 characters in length in TECH rather than ten as used in NONTECH. Each entry in a string of words is followed by an asterisk and its manually assigned word code number. Each string is 350 characters long, with a space separating each term/word code no. entry. Fourteen strings reside on one 2314 track. SWAMP contains all terms which are technical only when either preceded or followed by other technical or non-technical terms in an abstract. The grouping of words is identical to TECH, but in SWAMP, a HASH file reference number replaces TECH's word code number. String length and track accomodation is identical to that of TECH.

HASH contains numbered records or strings, the numbers corresponding to each term in the SWAMP file. Each string lists every technical use of the corresponding SWAMP term in conjunction with either its preceding or following word as found in the abstract. Each word combination listed is manually assigned a term code number similar to those found in TECH. An asterisk separates the word combination and its term code number while a space separates entries within a given string. Each string can accomodate 150 characters and 24 records occupy one 2314 track.

PROGRAMS: PATEX, TEC

PATEX reads each full text abstract, three words at a time, sorting the technical terms from the non-technical, printing unlisted terms, and then indexing the abstract by building, for each word in TECH and HASH, a threaded list of all the abstracts in which any given technical term appears. Each word is checked against three files, TECH, NONTECH, and SWAMP/HASH. If the word has no technical significance, it is ignored; if it conveys a technical concept, it is sent into the automatic indexing process. Should the term be absent in all three, it is printed for subsequent intellectual decision. Only one term is indexed at a time, but often the need to relate a word to its preceding and following words arises. PATEX gives consideration to this need in that it recognizes "bound" terms as having technical significance when in the company of certain modifiers. Thus three words are seemingly under scrutiny simultaneously. The indexing process converts every technical term (concept) to a term code number. MASTER file is searched for the appropriate term code number. Listed next to each such code number is the identification key number of the first abstract in which that word appears. A search of the list of that particular abstract's technical terms in ABSTECH file will reveal that the term corresponding to this word code number was indeed used in the abstract. The ten character number following this term code number in the ABSTECH list reveals the next abstract in which the term was used. A search of this second abstract's technical word list will turn up the same term code number and the identification key number of the third abstract in which the term was used is uncovered. The key number of the abstract presently being indexed is now inserted in the list of this last abstract on file. The abstract being indexed thus becomes the last entry in the threaded file.

TEC was written to make additions to the TECH file and to index those technical terms which could not be found on any of the Vocabulary/Thesaurus (V/T) files, therefore being subjected to human classification. However, its philosophy can easily be adapted to the maintenance of the other two files, NONTECH and SWAMP/HASH. As an abstract is being indexed by PATEX, a number of terms found in the abstract text may not be listed on any of the V/T files, due mainly to the vastness of language. These terms are printed and sent to the office for an intellectual decision on their technical value. Each word deemed technical in nature is assigned a

PROGRAMS: (Continued)

term code number and is read in either by card or on a terminal, with its corresponding term code number. The term-term code number entry is placed at the end of its appropriate TECH file word string, the proper string being determined by the word's length and its initial two letters. The term code number's position is checked on MASTER file in order to determine whether or not this word code number has already been assigned to the concept. An entry in this position indicates that the indexing of this concept term has taken place previously and so the ABSTECH threaded file is searched as it was in PATEX to insert this new term's abstract key number in the last position of the threaded file. If the MASTER word code number position is blank, the word code number is unique to this new term and the threaded file is initiated by placing the abstract number of the abstract in which this term was found into this blank area. Of course, the addition of the last identification key number along with the first identification key number, would permit the direct acquisition of the place to insert this new entry, thus precluding the necessity of threading through the file for each update time. The identification key number for the new entry would replace the identification key number that points to the end of the thread.

HARDWARE CONFIGURATION

The hardware configuration that was available for this project consisted of an IBM S/360 Model 40 computer operating under OS Extended Multi-Fixed task. The partition that was available was 31k, but 15k would have been more than adequate for the needs of this project. The terminal was a 2741 Communications Terminal and was interfaced with the CPU by a 2701 Control Unit. The connection was hardwire, unswitched line. Basic Telecommunications Access Method (BTAM) was used as the software package to access the terminal. A short control program was written in Basic Assembly Language to receive the messages from the terminal and to translate them into EBCDIC and to translate the CPU messages into 2741 code for reception at the terminal.

Programs originally written in PL/1 were stored as library load modules and could be called into core by the BAL program. Control was turned over to these load modules when they were loaded and input to them was either from the terminal or the card reader at the descretion of the programmer. Although no output was sent from these load modules to the terminal during the course of these experiments, they could have been by establishing the proper addresses through the Job Control language at load time. Programs written in any high level language could be compiled, linked and stored as library load modules for call by this BAL program, and data needed as input to them could be originated at the terminal, the card reader, or any other addressable device.

The 2741 terminal was selected for its compatibility to many task requirements at the University, the 2701 was selected for its immediate availability; it was to have been replaced by a 2702 or 2703.

BTAM was selected as the software package because of its relative simplicity and adequacy. PL/1 was selected as the programming language because of its many inherent string manipulation capabilities, and its overall versatility. BAL was used to write the calling and control program because it was the only language that could be used for this purpose.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
University of Dayton Research Institute Dayton, Ohio	UNCLASSIFIED
	2b. GROUP

3. REPORT TITLE DOCUMENT RETRIEVAL SYSTEM OPERATIONS INCLUDING THE USE OF MICROFICHE AND THE FORMULATION OF A COMPUTER AIDED INDEXING CONCEPT

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Final Summary Report, 1 December 1967 - 30 November 1968

5. AUTHOR(S) (Last name, first name, initial)

Scheffler, Frederic L.
Smith, Ralph B.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
December 1968	42	9

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
AF 33(615)-3389 b. PROJECT NO. 7381 c. Task No. 738103 d.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) AFML-TR-68-367

10. AVAILABILITY/LIMITATION NOTICES

This document has been approved for public release; its distribution is unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	AFML (MAAM) Wright-Patterson AFB, Ohio 45433

13. ABSTRACT Additional part-time indexers were trained according to a previously established training program. Many documents are now being received in microfiche form. Several makes of microfiche readers were evaluated, particularly with regard to their use by indexers. One make offered the most advantages, and a number of these readers were purchased. A method of producing typed hard copy abstracts from microfiche was found. The number of search requests increased with requests both from outside organizations and from the AFML accounting for the increase. The thesaurus was updated and a separate section containing only metallic materials terminology made the thesaurus easier to use.

A computer aided indexing concept was formulated which is based on analysis of technical text. Text words are matched against files of nontechnical words, technical words and bound terms. Words not recognized are presented for human intellectual decisions which are subsequently incorporated into the system. The concept included a self-contained document analysis, storage and retrieval system with which requestors could interact by means of remote access terminals. Remote communication capability with the computer was accomplished, some programs were prepared and file structures based on word length and alphabetization of the first two characters were designed.

This abstract has been approved for public release and sale; its distribution is unlimited.

DD FORM 1 JAN 64 1473

UNCLASSIFIED

Security Classification

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Text Analysis						
Information Retrieval Systems						
Document Retrieval Systems						
Indexing						
Automatic Indexing						
Computers						
IBM 360 Computers						
Computer Aided Indexing						
Remote Access Terminals						
Time Sharing Systems						
Computer Programs						
Microfiche						
Microfiche Readers						
Evaluation						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through

_____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through

_____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through

_____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

UNCLASSIFIED

Security Classification